Some Public Policy Problems with the Science of Carcinogen Risk Assessment[1]

Carl F. Cranor

University of California

Government agencies and private risk assessors use (quasi) scientific "risk assessment" procedures to try to estimate or predict risks to human health or the environment that might result from exposure to toxic substances in order to take steps to prevent such risks from arising or to eliminate the risks if they already exist[2] We might think of this as an aspect of "regulatory science".

Using carcinogen risk assessment as a model I consider several aspects of risk assessment—use of epidemiology, animal bioassays, and extrapolation models in predicting risks—to show how scientific procedures and uncertainties may determine (in a way often unbeknownst to the practitioners) the public policy debates concerning the estimation of risks from carcinogens.

For one thing, actual and possible scientific uncertainties are large enough that two different researchers using exactly the same data could come to quite different conclusions; such uncertainties permit risk assessors in large measure to determine regulatory decisions even though this exceeds their authority. In addition, both the uncertainties and science policies used to overcome them permit the infection of regulatory science with public policy, moral considerations if you will, thus making regulatory science much less like ordinary science.

Also, there is an assumption that risk assessment is independent of the risk management—of the important social policy—decisions and this seems mistaken. Because of the numerous uncertainties in risk assessment, scientific and nonscientific policy considerations bridge these gaps, infecting risk assessment with policy considerations. Thus, regulatory science, for the above and other reasons is much less like ordinary science than is normally supposed.

Furthermore, in human or animal statistical studies that are the foundation of standard setting scientists may be forced, for reasons of economics or circumstances, to study very small samples of experimental and control groups. This in turn in many circumstances forces either them or risk managers who use their data into a dilemma whether they recognize it or not: they are forced to choose between adhering to the evidentiary standards of good science as conventionally conceived—using cautious inferences before drawing scientific conclusions and avoiding false positives—and doing "good regulation"—providing results (estimates or predictions) protective of human health for purposes of regu-

---

PSA 1988, Volume 2, pp. 467-488

lation and avoiding false negatives. Because of these effects, I argue that scientists or policy makers (whichever is appropriate) should approach traditional avoidance of false positives as moral decisions appropriate to the context in question, especially when they are engaged in regulatory science.

Finally, the above results raise several questions scientists, philosophers of science, moral philosophers and policy makers should address forthrightly in order to serve better the aims of science and regulation.

1. Background

Risk assessment can be divided into toxicological and environmental risk assessment. Toxicological risk assessment is the assessment of risks to human health from particular substances making inferences from animal bioassays and human epidemiological studies, while environmental risk assessment estimates risks to human beings from exposure to carcinogens when they are released into the environment, into the soil, the ground water and air.[3]

Contrasted with *risk assessment* is *risk management*.[4] Risk assessment, a seemingly scientific enterprise, when used in regulatory law, aims at providing accurate information about risks to human beings so administrative agencies can regulate exposure to potentially cancerous substances in fulfillment of their respective statutory mandates. After scientists in the technical, scientific part of the federal agencies have provided an estimate of risks to human beings from exposure to toxic substances, they then give this information to the risk managers. *Risk management* is concerned then with managing the risks in accordance with statutory requirements and other economic, political and normative considerations.

At the outset, however, risk assessment in the present state of knowledge is a third best solution to the problem of estimating harms to human beings from exposure to toxic substances. The ideal is 'harm assessment'; if we had perfect information we would provide an accurate assessment of the harmful effects to people and the environment from exposure to toxic substances. This would provide us with exact numbers of deaths and diseases, thus we would not overestimate or underestimate the kinds and amounts of harms resulting from toxic exposures.

If we distinguish between risks and uncertainties, a "risk" is a *probability* of an unfortunate or undesirable outcome (Rescher 1983, p. 5), when one can assign such probabilities to outcomes. Thus, a "risk assessment" properly speaking aims at estimating the *probabilities* of harms from toxic exposures, and is a second best solution to the problem of estimating harms to human beings from toxic exposure.

At present the task of regulators is much worse than this, for great uncertainties obtain in trying to predict such harms. For example, some have argued that high dose to low dose extrapolation models used in animal bioassays can vary by six orders of magnitude (10^6), (Cothern, et. al. 1986) and many believe there is little biological basis for choosing between these models. (Freeman and Zeisel) Thus, we should think of risk assessments at present, not as risk assessments properly speaking, but as "risk and uncertainty assessments," the third best solution to identifying harm to human beings from toxic exposures.

In addition, matters of considerable moment depend upon the products of risk assessment, for in many cases one answer (a projection of high enough risks to require regulation) will impose considerable costs on the affected industry and perhaps the larger public, while another answer (a projection of risk low enough so that regulation is not a required) may well leave innocent people at risk from exposure to dangerous substances.

Because of substantial uncertainties, regulatory agencies run risks of making mistakes, risks of regulatory false positives and false negatives. A regulatory "false positive"

would occur when a substance is inappropriately regulated or regulated too stringently under a particular statute for the degree of harm that it causes. The substance might cause no harm at all, or much less harm than an agency believed it caused at the time of regulation. A regulatory "false negative" by contrast would be an outcome of a regulatory activity that resulted in a substance not being regulated at all when it should be, or regulated to a much lesser degree than it should be commensurate with the kind and degree of harm it causes and commensurate with the statute under which it was regulated. In a statistical study, because probabilities are involved, *it is certain* that if one were regulating large numbers of substances there would be both false positives and false negatives, for by chance alone, mistakes would be made. By analogy it is likely that agencies regulating large numbers of substances will make mistakes, either as a consequence of the underlying statistical studies or through other errors, which will result in both false positives and false negatives. Because of the possibility of regulatory mistakes it is a normative question concerning the parties on whom the costs of regulation or its absence should fall.

There are a number of ways to estimate risks to human beings from chemical substances—molecular structure analysis, short-term tests, long-term chronic bioassays in laboratory animals, and human epidemiology. In the following sections I summarize some results from the latter two to indicate some of the public policy problems that arise from the "science" in these fields.

2. Regulatory Science and Policy Choices

One method for estimating toxicological risks to human beings is to study the carcinogenic effects of substances on animals, and then to project risks to human beings based upon this information. An experimental group of rodents is fed high doses of a substance to see whether the cancer rate in the experimental group is significantly greater than the cancer rate in a control group. If it is, then scientists extrapolate from the high dose response rates in the rodents to project a low dose response rate in rodents (an exposure level much closer to the typical human exposure dose). Using this low dose response rate, based on principles of biology, toxicology and pharmacology (if such information is used), they then estimate, on the basis of rodent-to-human models, the likely risks which human beings would face at their levels of exposure. This risk information is then combined with exposure information in the workplace or in society at large in order to estimate the magnitude and extent of the risks to human beings. Finally, the risk information is combined with economic, policy, statutory, and technological feasibility information so that regulatory agencies can then decide how properly to manage the risks in question.

Such procedures require a number of inferences from the established experimental data from laboratory animals to the projection of end point risks to human beings, and these inferences have a number of uncertainties and gaps which must be bridged in order to produce the risk numbers. The uncertainties arise because there is insufficient information (in both theories and data) available to settle the scientific questions at issue. For example, the choice of different high dose to low dose extrapolation models (all of which have been advocated by respectable researchers) can produce variations in endpoint risk estimates that vary by a factor of 10^6. (Cothern, et. al. 1986) Use of different rodent to human extrapolation models can produce risk estimates that differ by a factor of thirteen. Use of different parameters in environmental fate models can produce results that vary by a factor of 500.[5] Some of these possible variations of estimates have not been exhibited in actual governmental risk assessments, for often different regulatory agencies have tended to agree on the same models, even though scientific data and theories are insufficient to support such choices. However, we have found in our research that actual risk assessments done by agencies may well differ by two orders of magnitude.[6]

In order to bridge the gaps created by uncertainties, a National Academy of Sciences Study has suggested that regulatory agencies adopt policies (inference guidelines) for

some fifty substantial "inferential gaps" that exist in the procedures for estimating risks to human beings from basic toxicological information. (National Research Council 1983)[7] These may not always be well grounded in biological fact.

These policies may govern decisions from choice of data, e.g., should toxicologists count benign tumors in experimental animals as providing evidence that a substance causes cancer, to choice of high dose to low dose extrapolation models, which aims at modelling statistically the mechanism of cancer, to choice of statistical procedures, e.g., whether one uses upper confidence limits or maximum likelihoods in the high dose to low dose extrapolation models, to whether certain kinds of data should be permitted or required to be offered in evidence in regulatory proceedings, e.g., whether agencies should consider pharmacokinetic mechanisms in calculating the dose to which a person is exposed.[8] Some of the main areas of disagreement between agencies and litigants to cancer standard setting are summarized in Appendix B.

The policies adopted to bridge such gaps can lead to considerable controversy, for they can make large differences in the estimation of risks to human beings, and there may be little or no biological plausibility to any of the models. (Freedman and Zeisel)

Thus, since science policies may have little justification in biological fact, they are chosen at least in part on public policy considerations. For example, most agencies have adopted the linearized multistage high dose to low dose extrapolation model in conjunction with a 95% upper confidence interpretation in order to avoid underestimating risks to human beings. (OTA 1987) The public policy outcome explicitly guides the choice of model. Thus, the "science" of risk assessment at the very outset, is widely infected with extra-scientific policy judgments about the matters at issue, and the "science" of risk assessment is less than fully scientific in the ordinary sense.

If the above view is correct this poses a dilemma for risk assessment science. On the one hand, an attempt to make it more scientific in the sense of basing component models on established biological facts will result in little regulation because the knowledge is not available and because current regulatory (and tort) laws tend to preserve the status quo until evidence for changing it is provided. On the other hand, if agencies are to expedite risk assessments and not wait for answers to these scientific questions, then the science of risk assessment will be substantially permeated with nonscientific policies and thus be quite different from ordinary science.

A larger point is that the many uncertainties which at present exist in cancer risk assessment may make it difficult, if not impossible, for scientists to remain wholly faithful to scientific traditions while providing data which will permit timely and morally justifiable regulations. More likely, fidelity to scientific tradition will paralyze regulatory activity. In addition, however, even if there were not the above problems, there remain more fundamental problems with aspects of carcinogen risk assessment that should be of concern to scientists, philosophers of science and policy makers. These are considered in the remainder of the paper.

3. Hidden Policy Tradeoffs

(A) I have argued elsewhere that a wise and conscientious epidemiologist (or a risk manager using the epidemiologist's results) doing a cohort epidemiological study (OTA 1981, p. 137) with perfect evidence, but with constrained sample sizes for the disease being studied, *cannot* but face potentially controversial moral and social policy decisions in order to design and interpret an epidemiological study and to produce the risk numbers that are the outcome of such work. I summarize some of the main points of that argument here.[9]

Consider the theory of hypothesis acceptance and rejection. This provides appropriate terminology to characterize the main risk and proof variables with which epidemiologists must work, and to understand the logic of scientific proof available in this area.

Consider for regulatory purposes whether benzene causes serious disease, e.g. leukemia or aplastic anemia. The null hypothesis would predicate, e.g., that exposure to benzene is *not* associated with greater incidence of a certain disease than that found in a nonexposed population, while the alternative hypothesis would then indicate that exposure to benzene *is* associated with a greater incidence of such diseases. (Feinstein 1977, pp. 320-321)

By chance alone a researcher investigating these hypotheses risks false positives [type I errors], or false negatives [type II errors]. The probability of committing a type I error is normally designated **A** and the probability of committing a type II error is designated **B**.[10] (Feinstein 1977, pp. 320-321) The "power" of a statistical test is 1 - **B**. Conventionally, A is set at .05 so that there is only a one in twenty chance of rejecting the null hypothesis when it is true. (Walter 1977, p. 391)[11]

The low value for **A** probably reflects a *philosophical view* about scientific progress and may constitute part of its justification.[12] In building the edifice of science, by keeping the odds of false positives low, one ensures that each brick of knowledge added to the structure is solid and well-cemented to existing bricks of knowledge. Were one to tolerate higher risks of false positives, take greater chances of new knowledge being false by chance alone, the edifice would be much less secure. A secure edifice of science, however, is not the only important social value at stake in environmental health regulation.

One can think of **A**, **B**, and 1 - **B** as measures of the "risk of error" or "standards of proof." What chance of error is a researcher willing to take? Is a twenty percent (**B** = .20) chance of a study showing benzene does not cause cancer, when in fact it does, an acceptable risk? When workers or the general public may be contracting cancer (unbeknownst to all) even though a study (with one kind of high epistemic probability) shows they are not, is a risk to their good health worth a twenty percent gamble?

Alternatively, we might think of **A**, **B**, and 1 - **B** as standards of proof. How much proof do we demand of researchers and for what purposes? Must potential carcinogens be condemned by mere preponderance of evidence, somewhat more than fifty percent of the evidence (e. 1 - **B** = .51+)? That is to say, must researchers be more than fifty percent sure that benzene is a carcinogen presenting a risk to employees in the workplace before regulating it? Should scientists in agencies be permitted to take a forty-nine percent chance (**B** = .49) that substances are not high risk carcinogens to the populace, when in fact they might be?

These trade-offs are also a function of two other variables: N, the total sample of people in the exposed and unexposed groups, and **D**, the relative risk one wants to detect. (Feinstein 1977, pp. 320-324) With **A** and **B** fixed, the relative risk one can detect is inversely related to sample size: the smaller the risk to be detected the larger the sample must be. The value of **D** for which a study might be designed to detect depends upon many factors, including the seriousness of the disease, its incidence in the general population, and how great a risk, if any, the exposed group justifiably should be expected to run.

Furthermore, **A**, **B**, **D** and N are mathematically interrelated. If any three of them are known the fourth can be determined. Because the variables are interdependent, crucial trade-offs may be forced by the logic of the statistical relations. Consider the following hypothetical decision tree (summarized in Table I) which presents five related alternatives.

The first thing to notice is that if one wished to have the most *accurate* study with equally small chances of false positives and false negatives, one should adopt alternative

(1). This choice is likely to be very expensive or such large samples will be unavailable for study. Similar conclusions follow for alternative (2).

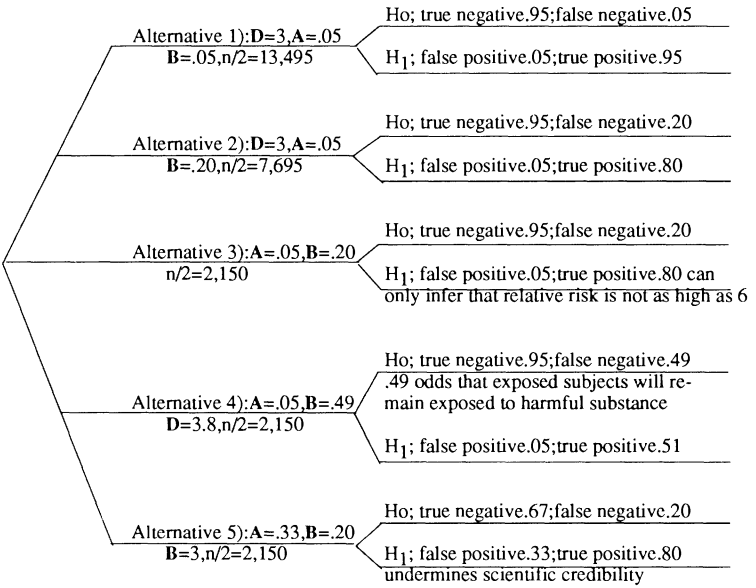
If one rejected these alternatives, however, and were forced to study a much smaller sample, this poses a dilemma. Alternatives (3) and (4) may leave those exposed to toxic substances at considerable risk, because by chance alone they run substantial risks of false negatives (or of failing to detect a relative risk that is of concern). Alternative (5) risks undermining the credibility of the research because it is inconsistent with scientific practice. Thus, the mathematics of epidemiology together with small sample sizes and a low background disease rate impose difficult moral choices on researchers and regulators alike when such studies are used in regulatory contexts to estimate risks to people.

Table I

Cohort Study

Assume that the prevalence of disease L in the general population is 8/10,000.

The study seeks to detect a relative risk of 3 (**D**=3), provided such risk exists.1



Nearly all the figures used in these examples are taken from my "Epidemiology and Procedureal Safeguards for Workplace Health in the Aftermath of the Benzene Decision," which figures were originally supplied to me by Dr. Helaine Pleet and the computers at the Centers for Disease Control in Atlanta, Georgia. That information is summarized in Appendix A. Similar numbers can be calculated from the equation (4) Walter (1977) in note 47 infra. The particular numbers in alternative (1) come from line 3 (c) of Appendix E.

Alternative (3) suggests some interesting results for "negative" or "no effect" studies. Assume a study is run on 2,150 exposed workers with **A** at .05 and **B** at .20, when the prevalence of the underlying disease is 8/10,000. With these values, we only could be

confident of detecting a relative risk of six with a power of .80. But suppose none were detected, that is, the study was "negative" or showed "no effect" between some chemical C and the disease L. What could we infer? At most we would be justified in concluding that the relative risk was less than 6. It might be 5.9 or 1, but given the constraints on the study, we could not conclude so statistically. Thus, for "no effect" studies the most that can be inferred is that the relative risk to people in the exposed group is not as high as the relative risk tested for in the study. Regulatory agencies regard such results as useful mainly for setting upper bounds on risks to people. (OTA 1987)

Furthermore, negative studies based on small samples and low **A** can have quite unacceptable type II errors for public policy purposes. Needleman and Bellinger, surveying fourteen epidemiological studies of lead in children, with low **A** (< .05) found that type II error rates ran as high as 82%. Thus, by chance alone researchers had up to 82% odds of failing to detect adverse effects on children from lead exposure even when they may in fact have existed. (Needleman and Bellinger)[13]

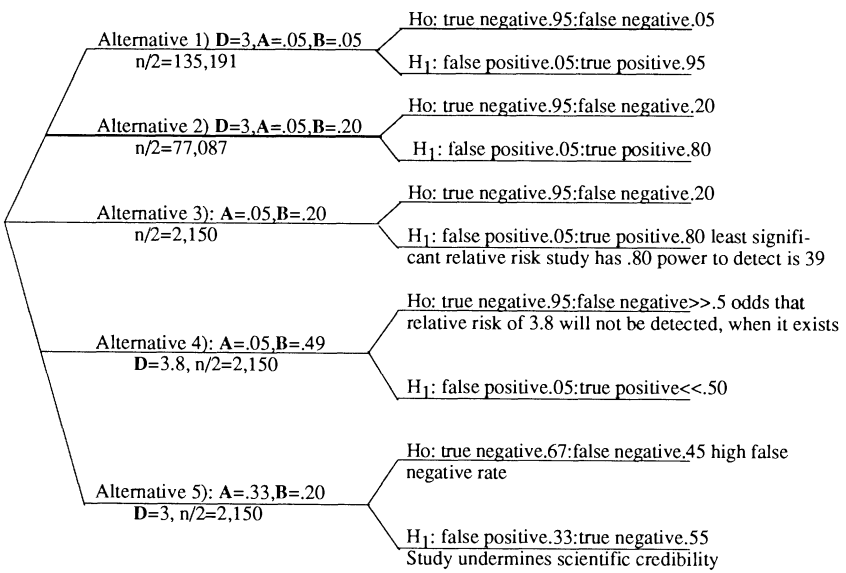
In other cases it may be statistically impossible to detect a significant health risk when one exists. Suppose that it is thought important to detect a relative risk of three among workers exposed to toxic substances for a disease that occurs in eight people of every 10,000. If there were only 1,000 workers to study (with **A** at ,05 and **B** at .20), a relative risk could not be detected below 10 with a power of .80, even if it turned out the substance in question caused a threefold increase in mortality. (Walter 1977, p. 39)

Table II

Cohort Study

Assume that the prevalence of disease L in the general population is 8/100,000.

The study seeks to detect a relative risk of 3 (D=3), provided such risk exists.



As striking as the preceding examples are, they only suggest the statistical problems a cohort study of a typical environmentally caused disease (e.g., benzene-induced leukemia) might pose. If the prevalence of the disease subject to study were rarer by a factor of 10, a more realistic number for leukemia, (National Cancer Institute 1981, pp. 662-63, Table 51) then the decision tree exhibits even more surprising results (summarized in Table II). In an analogue to alternative (2), holding **A**, **B** and **D** constant and changing the incidence of L to 8/100,000 for leukemia, one would have to study 77,087 people exposed to benzene and an equal number not exposed to obtain statistically significant results.[14] In an analogue to alternative (3), one would have to study 21,580 subjects, one could detect a relative risk of six.[15] Forced to study 2,150 subjects, one could detect a statistically significant relative risk no lower than thirty-nine.[16] If one could only study as few as 1,000 people, one could reliably detect a relative risk no lower than fifty.[17] Studying as many as 10,000, one could detect a relative risk of no less than nine.[18] In alternative (4), holding **A**, **D** and **N**/2 constant, one would lack even fifty percent confidence in one's results.[19] Even increasing **A** to 16% would not decrease the chance of a type II error below fifty percent.[20] Analogues of alternative 5) lead to similarly unsatisfactory results. The point: the rarer a disease, the greater the problems faced by epidemiologists (or policy makers interpreting a study).

The point of the above examples is that epidemiologists even with no evidence gathering problems, when faced with the relatively rare diseases and small samples available for study typical of environmentally induced diseases, must take into account cost considerations, samples available for study, and some of the objectives of the study in order to *design* and *conduct* the study in question. The most important of these questions concerns the odds of false positives and false negatives researchers are willing to take. In regulatory contexts this will make a difference on whom the costs of possible statistical mistakes will fall. *Thus someone*, either the *scientists* designing and conducting the study or the *policy makers* relying on the statistical evidence and applying the study results in a regulatory context, must face up to some crucial social issues. Who should bear the costs of possible statistical mistakes: The company making the substance which may be the object of regulation or the public or employees of a firm who might be subject to health harms, if the substance is toxic?

Furthermore, if, as a matter of scientific practice scientists uncritically remain committed to the 95% rule (A = .05) and forward the results uncritically to risk managers, they risk deciding the public policy question of whether a risk to human health exists or not merely by how they present the evidence. *Uncritical* commitment to the 95% rule convention may beg the public policy question at issue. Whether or not this occurs depends upon the practices of the scientists involved; this is a matter of sociology. It is possible for statisticians to present their results objectively without commitment to preset or automatic **A** and **B** values,[21] but then the **A**, **B** value tradeoff is merely pushed back one step to the risk managers.

The statistics of animal bioassays exhibit behavior similar to that of epidemiology although the numbers are not quite as dramatic. Talbot Page has shown that if one has fifty control animals and five of these develop tumors at one site, e.g., the liver, while twelve of fifty treated animals develop tumors at that same site, reliance on the 95% rule would reject this as statistically significant evidence of a difference in tumor rates.[22] Nonetheless, use of Bayes' Theorem and some plausible background assumptions would show the tumor rate in the treated animals compared to the controls to be a matter of considerable concern.[23]

Just as in the case of epidemiological studies, in animal studies the rarer the disease rate is in control animals and the fewer the controls in the sample with tumors, the more researchers should consider using higher A values to ensure that existing diseases in animals do not go undetected because of the scientific conventions of the statistics of the studies.[24]

(B) One can see additional problems with understanding epidemiological studies for regulatory use by looking at how fixed data might be interpreted once a study has been finished. The fixed data would consist of the background disease rate, sample size, and observed relative risks. However, for purposes of *interpreting* this information epidemiologists or policy makers could vary the values of **A** and **B**. Consider one possibility.

Suppose the study of 2,150 exposed individuals produced an observed relative risk of about three, because there were five deaths compared with 1.72 (one or two) in the control group. Is this a positive result or not? The following table shows that one could interpret the study as a positive or as a negative study for any of several pairwise choices of **A** and **B** values.

Positive Results			**Negative Results**
A	**B**	**D** (Least significant relative risk which test has a power of .51 or higher to detect)	
.10	.49	3.0	When A<.10 (with B constant) or B<.49 (with A constant)
.15	.40	3.0	When A<.15 (with B constant) or B<.40 (with A constant)
.20	.30	3.1.	When A<.20 (with B constant) or B<.30 (with A constant)
.25	.25	3.1	When A<.25 (with B constant) or B<.25 (with A constant)

Any pairwise combinations of **A** and **B** in the left hand column will show that the study outcome is positive. Changing the variables slightly as indicated in the right column will produce a negative study.[25] This example shows that risk assessors or policy makers have considerable flexibility in *interpreting* the data of a study. How they *interpret* and *use* the evidence in certain regulatory and legal contexts will have important consequences for protecting human health. Similar results can be reproduced for interpreting the results of animal bioassays; researchers and policy makers should be given advice similar to that given for interpreting epidemiological studies: adherence to low **A** values (e.g., use of the 95% rule) may well bias the normative, regulatory issues in unfortunate ways. Thus, the interpretation of the tumor rates of animal bioassays should be approached as a normative, or moral question.

(C) An additional problem concerns the relative values of **A** and **B** when evaluating the health effects of relatively large numbers of substances. As long as **A<B** and **A** is in the neighborhood of .05, we are doing "better" science conventionally conceived, but we are also protecting possibly harmful chemicals better than human health. Suppose that we have twenty-four hundred substances to test. Assume also, that 40% of those are carcinogens and 36% of them are not, with the remainder equivocal or inconclusive (realistic numbers according to an OSHA study). Now if epidemiologists set **A** at .05 and **B** at .20 (typical values), assuming this is large enough sample, we will have one-hundred and ninety-two false negatives and forty-three false positives. With one-hundred and ninety-two false negatives, this means that one-hundred and ninety-two substances will pose some risk of cancer to the populace (and how large a risk this is will depend upon the prevalence of the disease, the

relative risk associated with the substance, the substance's potency and the number of people exposed), but the test will not show it. On the other hand, forty-three false positives mean that forty-three substances will be wrongly regulated (or possibly banned altogether), depending upon the statutory authority in question. If substances are banned, the products into which they are incorporated will be more expensive to produce and market, or we will be deprived of their use and benefits altogether. If the substances are merely regulated, they likely will be more expensive to produce and market.

4. Philosophical Decisions in Regulatory Science

In certain common circumstances, described above, in which we use or need to use the tool of epidemiology, or other statistical studies in estimating risks to human beings from toxic substances, and in which scientific tradition would ordinarily rely upon low **A** values, e.g., **A** = .05, the study of relatively rare events by means of small samples forces a tension between the use of this rule and other public policy and moral concerns we might have. Roughly the tension is between a commitment to traditional scientific caution in pursuit of the truth (represented in the 95% rule), and a commitment to protecting people's health, or at least not taking chances with their health. However, the same examples show that in the circumstances described there is no necessity to the received scientific practice—it could be done differently. *Whether epidemiologists or policy makers should adhere to the 95% rule in certain contexts is a normative, a philosophical question.* Moral philosophers, philosophers of science, lawyers and those in public institutions with the authority to protect our health should explicitly acknowledge and address these issues. Those charged with regulating our exposure to toxic substances should consider such policy problems in the design of the study.

Second, the use of epidemiological data is not obviously a neutral and objective project. In subsection 3(B) above, sample size and number of deaths are fixed data in the study, but *whether a risk to human health is judged to exist depends upon the choice of values for* **A** *and* **B**. How the fixed data gets used in subsequent regulatory or legal proceedings will also depend upon these variables and may have important consequences for our health.

More importantly, in order to design and perform the studies in question, the choice of values for **A** and B commits scientists or the policy makers who use the studies implicitly, if not explicitly, to making judgments that are the equivalent of moral or social policy considerations.

Thus, since the design, reporting, interpretation and use of epidemiological data are dependent upon such judgments, and since we could change traditional practices for interpreting and using scientific evidence, we should face the use of the 95% rule in risk assessment and risk management proceedings as a normative question.

Consider an analogy in the law. In criminal trials, avoiding wrongful damage to someone's reputation and well-being is so important that we spend considerable sums of money and deliberately make proof of guilt very difficult in order to avoid wrongly inflicting harsh treatment and condemnation on the defendant. We could save money and have more unjust outcomes if we thought it worth the human costs, but we do not. Clearly a number of moral and cost considerations have influenced the institution of the criminal law. With regard to criminal trials, we have been quite self-conscious in debating the moral considerations that bear on the design and workings of such institutions. Similar problems arise in the interpretation and use of epidemiological studies, and I am suggesting that similar debates should attend the use of scientific results that may have such profound influences on our health.

A further problem is that the conventional evidentiary practices of science (use of the 95% rule) may well be much more demanding than the evidentiary requirements and aims

of the tort and regulatory law. Tort law requires a plaintiff to establish his case by a "preponderance of the evidence." If this standard were expressed in quantitative terms it would require that somewhat more than 50-55% of the evidence favor plaintiff's case. In regulatory proceedings for the most part a lesser evidentiary standard is required than in the tort law to establish a case as long as the agency does not act arbitrarily or capriciously.

Thus the problem: the evidentiary standards of science as exhibited in the 95% rule are much more demanding that the legal standards of evidence where the regulatory science evidence will be used. Thus by default if regulators use conventional scientific standards their science will in many cases beg the regulatory question against regulation. Using epidemiology, as an example, when relatively rare diseases and small samples force regulatory scientists (or policy makers) to choose between high false negatives or high false positive rates that would be intolerable for normal scientific work, if they choose to tolerate high false negatives and protect the integrity of their scientific work, they thereby favor non-regulation or less regulation by their choice of evidentiary standard.

Furthermore, given the wider aims of both the regulatory and tort law and the weaker evidentiary standards that must be met in the law, there may not be good reasons *in these legal contexts* to require regulatory science to meet the same evidentiary burdens as normal scientific work aimed only at discovering the truth and adding to our stock of scientific knowledge about the world. Thus, I would urge that for legal purposes regulatory science adopt evidentiary standards much closer to that of the legal institutions it is meant to serve (although this argument must be prosecuted elsewhere).

5. Conclusions

Several conclusions emerge from the arguments presented above. First, because of substantial uncertainties in the risk assessment process, in nearly all cases it is difficult, if not impossible, to obtain *scientifically* respectable and accurate estimates of health risk to human beings. A number of the crucial inferences at present are not grounded in biological fact.

Second, in order to overcome the absence of scientific information, regulatory agencies have had to resort to extra - scientific policy judgments, resulting in subversion of the policy neutral status of regulatory science as it is presently practiced. Thus, the risk assessment policies as well as particular assessments of the risks posed by particular substances are not objective in the sense in which mechanics or electromagnetism is thought to be objective, and free from moral considerations.[26]

Third, and most important, however, is that even in circumstances in which there are no problems with scientific uncertainties or with extra-scientific policy considerations infecting regulatory science, substantial nonscientific policy judgments enter into interpreting and using the scientific evidence. Further, researchers or risk managers in interpreting and reporting the results of epidemiological studies face the equivalent of social policy decisions. Choice of **A** and **B** variables will influence how the study gets reported and may subsequently affect the regulation of toxic substances.

I emphasize these results in order to occasion debate among scientists, philosophers of science, moral philosophers, and lawyers concerning the best set of procedures to adopt in the face of such issues. It raises at least the following questions: Should the intellectual and institutional integrity of the relevant areas of science remain free from the nonscientific or extra-scientific policy judgments, if doing so would mean no regulation or would substantially slow the risk assessment process against regulation of toxic substances? Since scientific truth is not the only social goal in risk assessment, to what extent should pursuit of other aims modify that goal; in particular to what extent should the protective health aims of regulation or the need for expeditious risk assessments modify or shape pursuit of scientific truth? To what extent should the "institution" or practices of science

accommodate the demands of other institutions in our society, e.g., regulatory institutions as well as the tort law?

I have only provided the background for posing these larger questions, but they should be faced by appropriate parties interested in the interface of science, law and public policy concerns in regulatory science. Although I have not answered such questions in this paper, they are being pursued in forthcoming papers. (Meetings/PSA 03/21/89)

Notes

[1]Research on the paper has been supported by a grant from the University of California Toxic Substances Research and Teaching Program for the UC Riverside Carcinogen Risk Assessment Project of which the author is the principle investigator. I am indebted to Deborah Mayo, D.V. Gokhale, and William Kemple and Kenneth Dickey for comments and criticisms on the ideas presented in this paper.

[2]When risk assessment is used in toxic tort suits, the aim is to provide evidence of harm to provide a basis compensation for injuries caused by such substances.

[3]This distinction is adapted from work done by the University of California, Riverside, Carcinogen Risk Assessment Group.

[4]The National Research Council. (1983), *Risk Assessment in the Federal Government: Managing the Process*. Washington D.C.: U.S. Gov't Printing Office pp. 18-19.

> We use *risk assessment* to mean *the characterization of the potential adverse health effects of human exposures to environmental hazards*. Risk assessments include several elements: 1) *description of the potential adverse health effects* based on an evaluation of results of epidemiologic, clinical, toxicologic, and environmental research; 2) *extrapolation from those results to* predict the type and estimate the extent of health effects in humans under given conditions of exposure; judgments as to the 3) *number of characteristics of persons* exposed at various intensities and durations; and 4) *summary judgments* on the existence and overall magnitude of the public-health problem. Risk assessment also includes characterization of the uncertainties inherent in the process of inferring risk...
>
> The Committee uses the term *risk management* to describe the *process of evaluating alternative regulatory actions and selecting among them*. Risk management, which is carried out by regulatory agencies under various legislative mandates, is an agency decision-making process that entails consideration of political, social, economic, and engineering information with risk-related information to develop, analyze, and compare regulatory options and to select the appropriate regulatory response to a potential chronic health hazard. The selection process necessarily requires the use of value judgments on such issues as the acceptability of risk and the reasonableness of the costs of control.

[5]Research done under the UCR Carcinogen Risk Assessment Project shows that the predicted movement of benzene escaping from leaking gasoline tanks can vary by a factor of 500. (Lee and Chang)

[6]See appendix **A** for a summary of several risk assessments done on each of benzene and ethylene dibromide and their differing conclusions. These results were compiled by Kenneth Dickey.

[7]Agencies have adopted four kinds of assumptions:
1. assumptions used when data are not available in a particular case;
2. assumptions potentially testable, but not yet tested;
3. assumptions that probably cannot be tested because of experimental limitations; and
4. assumptions that cannot be tested because of ethical considerations.
(The Office of Technology Assessment 1987).

[8]If the information were available and understanding of mechanisms were adequate (both of which appear to be doubtful at present for most substances), this should enable scientists to distinguish between the dose administered to a rodent or person external to the body from the dose which reaches the target organ where it would then do some damage.

[9]The conclusions of that earlier paper have been modified somewhat because of arguments presented by Deborah Mayo, a cosymposiast at the PSA meetings where this paper was read. See (Mayo 1989).

[10]Throughout I use the bold face **A**, **B** & **D** for Greek alphabet letters Alpha, Beta and Delta because the computer lacks Greek capabilities.

[11]This "conventional" value varies even within research science. For some fields a 5% chance of false positives is unacceptably *high*, e.g., in microbiology or in subatomic physics. For regulatory purposes, however, I suggest (below) this value may be unacceptably *low*.

[12]The low value for **A** may also be a mathematical artifice explained historically. As Giere puts it "The reason [for the practice of having a 95% confidence level to guard against false positives] has something to do with the purely historical fact that the first probability distribution that was studied extensively was the normal distribution." (Giere 1981, pp. 212-213) Two standard deviations on either side of the mean of a normal distribution encompasses 95 percent of the entire distribution.

He adds "95 percent is a comfortably high probability to take as the standard for a good inductive argument. Most scientists seem to think that science can get along with one mistake in 20, but not with too many more." (p. 213)

[13]Needleman's and Bellinger's results of type II errors are summarized in Appendix C.

[14]This figure is from line 3(d) of Appendix E.

[15]This figure is from line 5(d) of Appendix E.

[16]This figure is from Walter 1977, p. 388 (equation (4)).

[17]This figure is from Walter 1977, p. 391 (Table 2).

[18]Id.

[19]I put this point in a general way because of statistical problems. A sample of 2,150 people is so small compared to the disease rate of leukemia that the assumptions underlying the usual epidemiological equations no longer apply. The most important of these assumptions is that there is a normal distribution of diseased individuals through the population. In a group of 2,150 people on the average only .17 people would contract leukemia. Since people come in multiples of one, a probability procedure is needed to estimate how many times a coin will come up heads if it is flipped twenty-five times. The chance of k individuals contracting leukemia (with a prevalence of 8/100,000) is given by the formula for a Poisson distribution:

$$p(k)=e^{-M}\frac{(M)^K}{K!}$$

where e is the mathematical constant and M is the mean of the sample population being studied, when A=l.5%, D=3, and **B**=90.5%. Thus, the chances of making a type II error are nearly one hundred percent. When A=l6% and **D**=3, **B**=59.7%, one could not be more than forty percent confident that one could detect a relative risk of three, even when one existed. Professor David J. Strauss of the Department of Statistics, University of California at Riverside, pointed this out to me.

[20]Id.

[21]See (Mayo 1989) who argues this point.

[22]Page, T. 18.

[23]Our research shows Page is seriously mistaken about some aspects of hypothesis testing and use of the 95% rule, for the power of appropriate tests in his case is quite high, even though the test is not statistically significant by the 95% rule. The power for seventeen control and treatment tumors when A=.0542 is about .9468, while the power of the test for sixteen treatment and control tumors is .9630 when A=.0857. This point was established by the author and Bill Kemple working on the UCR Carcinogen Risk Assessment Project.

[24]Results of representative animal bioassays are taken from research under the UCR Carcinogen Risk Assessment Project by Mr. Bill Kemple and the author, and summarized in Appendix C.

[25]The data for this table are from Appendix F.

[26]In addition, particular science policy judgments may favor one set of parties to regulation over another. For example, several of the science policies are chosen to avoid false negative regulatory mistakes, thus risks to human beings tend to be overestimated leading to greater rather than lesser regulation of the substance in question. Risk assessment policies, however, might be designed to be more nearly procedurally neutral in that if the science policies have all been adopted after sufficient opportunity for notice and comment and in accordance with the Administrative Procedure Act, all the parties to the regulation are playing by the same set of rules.

Appendix A

Estimates of Lifetime Excess Cancer Risk from Exposure to Ethylene Dibromide

Risk Assessment	Excess deaths per 1,000 (95 pct upper confidence limits)		Model
	20 PPM	0.1 PPM	
EPA CAG:			
1976.....	999	67	"Upper bound" on onehit
1980.....	999	45	Onehit.
SRI.....	990-1000	117	Multihit, onehit.
OSHA In-House....	160-437(251-516)	0.06-3(1.4-3.6)	Onehit, multistage.
Busch.....	190-490(293-588)	.....	Probit.
Brown:			
Hemangiosarcoma.	70-110(134-148)	0.2-0.6(0.7-0.8)	Onehit, multistage.
Nasal Tumors....	725(785)	6 (8)	Onehit.
CAL/OSHA.....	400-996	0.5-2	Onehit, linear.

Benzene Risk Estimates by Agency (a)

10 ppm	EPA 34	OSHA 95	White 44-152	NIOSH 634	IARC 14-140	API 8
1 ppm	3.4	10	5-16	5	1.4-14	0.6

(a) Lifetime cancer per 1000 people exposed.

Compiled by Kenneth Dickey, graduate student in Philosophy and Research Assistant in the UCR Carcinogen Risk Assessment Project.

Appendix B

Decision Points in Quantitative Risk Assessment:
Alternative Data Choices and their Impact on Risk Estimates

Issue	More safety-oriented risk estimate	Less safety-oriented risk estimate
Input Data		
1. "Benign" tumors*	Data used	Data not used
2. Exposure (latent period)*	Allowance for "wasted exposure"	No allowance
3. Use of epidemiologic data*	Not used	Used, or suggestion that one wait for, epidemiologic data
4. Controls in epidemiology*	Detailed attempts	National data: (at times) "locally adjusted" controls
5. Risk groups in epidemiology*	Narrow - clearly exposed	Broad - doubts about exposure
6. Exposure estimate (actual population)	Upper 5% (or 1%) of population	Average in population
7. Use of "negative" studies	Not used	1. Used 2. Equated with positive studies (1=1)
8. NOGL (LOGL, etc.)	Concern with the experimental sample size	No attention is devoted to sample size
9. Adjustment for background exposures	Additivity in dose	Additivity in the response. Abbott's correction
10. Species (animal) extrapolation	"most sensitive"	Some average (usually geometric mean)

Issue	More safety-oriented risk estimate	Less safety-oriented risk estimate
Model		
1. Threshold*	No	Yes
2. Model	One hit	"Sensitivity" model (eg., Probit/Logit)
3. Concern with differentiating "genotoxicity" in animal bioassays*	No	Yes
4. Species (animal) for extrapolation	"Most sensitive"	Some average (usually geometric mean)
5. NOGL (LOGL, etc.)	Concern with the sample size	No attention is devoted to sample size
6. Adjustment for background response	Additivity in dose	Additivity in the response. Abbott's correction
Interpretation of Results		
1. Safety factor	Large, eg., 1000 or more	Small, eg., 10-100
2. Who is to be protected?	"Sensitive" (eg., pregnant women, children, old people)	Average
3. False positives/ false negatives	"Cost" of false negative is greater	"Cost" of false positive is greater
4. Concern for multiple sources of exposure	Yes	No
5. "Acceptable" risk	1 x 10-6 or less (lifetime)	Greater than 1×10^{-6} (lifetime; eg., 1×10^{-6} per year)

*For computations on carcinogens.

Source: Adapted from Table: Issues and assumptions in risk assessment computations, in Schneiderman, M., "Quantitation and Interpretation in toxicology: What Can We Do with the Numbers?", 1986, unpublished.

Appendix C

Type II Fallacies in the Study of Childhood Exposure to Lead at Low Dose: A Critical and Quantitative Review

Herbert L. Needleman
David C. Bellinger

Amended Metanalysis - 9/30/86

Author	Year	N	Effect Size	Power Small Effect*	PBAL(IT)	2LOGeP
Ernhart Et	1974	80	0.6	0.2	0.025	7.38
Needleman	1979	73	0.35	0.47	0.015	8.4
Yule, et.al.	1981	82	0.573	0.42	0.021	7.73
Winneke, et.	1982	26	0.26	0.18	0.15	3.7
Smith, et.al.	1983	185	0.17	0.7	0.12	4.24
Winneke Et	1983	115	0.351	0.25	0.4	1.83
Harvey	1984	47		0		
Shapiro	1984	193	0.46	0.48	0.025	7.38
Lansdown	1986	162	0.07	0.48	0.66	0.83
Hansen	1985	82	0.5	0.34	0.0005	15.2
Hawk Schro	1985	75	0.64	0.25	0.0004	15.64
Schroeder	1985	104	0.5	0.33	0.005	10.6
Fulton	1986	501	0.4	0.52	0.003	11.6
Hatzakis	1986	509	0.4	0.52	0.00065	14.6

dx = 109.13
df = 26
P = 3×10^{-12}

* Type II error rates are 1- (the number in the column "Power of small effect"). Thus, the Winneke, et.al. study had a Type II error rate of .82.

Appendix D

Representative Power Values for Fixed Alphas and Disease Rates in Groups of Fifty Control and Fifty Treatment Animals In Animal Bioassays Using Fisher's Exact Test*

(Type I Error)	P1 (Tumor Rate In Control Group Of Fifty Animals)	P2 (Tumor Rate In Treatment Groups Of Fifty Animals)	Power of Test
.05	.01	.03	.0104
	.025	.075	.126
	.05 (2 or 3 animals)	.15 (7 or 8 animals)	.3878
	.075 (3 or 4 animals)	.225 (10 animals)	.5927
	.1	.3	.7469
.10	.01	.03	.0391
	.025 (1 animal)	.075 (3 or 4 animals)	.2295
	.05 (2 or 3 animals)	.15 (7 or 8 animals)	.542
	.075	.225	.7359
	.1	.3	.8535
.15	.01	.03	.1195
	.025	.075	.3646
	.05	.15	.6275
	.075	.225	.8017
	.1	.3	.8995
.20	.01	.03	.1338
	.025	.075	.4426
	.05	.15	.7042
	.075	.225	.8537
	.1	.3	.9290
.25	.01 (0 or 1 animals)	.03 (1 or 2 animals)	.2894
	.025 (1 animal)	.075 (3 or 4 animals)	.5314
	.05	.15	.7672
	.075	.225	.883
	.1	.3	.9532
.3	.01	.03	(?).2894
	.025	.075	.5328
	.05	.15	.8024
	.075	.225	.9212
	.1	.3	.9641

*Compiled by Bill Kenple, graduate student, Department of Statistics, and Research Assistant, University of California, Riverside, Carcinogen Risk Assessment Project.

Appendix E

Representative Statistical Values for Prospective Epidemiological Studues of Diseases at Various Incidence Rates

Incidence of disease in general population:		Numbers of Subjects to be Studied N/2 exposed	N/2 unexposed	N total
1) δ = 1/5 (relative risk) α = .05 (risk of type I error) β = .20 (risk of type II error)				
a) 8/100	a)	691	691	1,382
b) 8/1,000	b)	7,631	7,631	15,262
c) 8/10,000	c)	77,022	77,022	154,044
d) 8/100,000	d)	770,939	770,939	1,541,878
2) δ = 1.5 α = .05 β = .05				
a) 8/100	a)	1,212	1,212	2,424
b) 8/1,000	b)	13,382	13,382	26,764
c) 8/10,000	c)	135,078	135,078	270,156
d) 8/100,000	d)	1,332,037	1,332,037	2,664,074
3) δ = 3 α = .05 β = .20				
a) 8/100	a)	62	62	124
b) 8/1,000	b)	756	756	1,512
c) 8/10,000	c)	7,695	7,695	15,390
d) 8/1000,000	d)	77,087	77,087	154,174
4) δ = 3 α = .05 β = .05				
a) 8/100	a)	109	109	218
b) 8/1,000	b)	1,326	1,326	2,652
c) 8/10,000	c)	13,495	13,495	26,990
d) 8/100,000	d)	135,191	135,191	270,382
5) δ = 6 α - .05 β = .20				
a) 8/100	a)	13	13	26
b) 8/1,000	b)	207	207	414
c) 8/10,000	c)	2,150	2,150	4,300
d) 8/100,000	d)	21,580	21,580	43,160
6) δ = 6 α = .05 β = .05				
a) 8/100	a)	22	22	44
b) 8/1,000	b)	363	363	726
c) 8/10,000	c)	3,771	3,771	7,542
d) 8/100,000	d)	37,845	37,845	75,690

From Cranor (1983), pp. 400-401

Appendix F

Relative Risk as a Function of Alpha Beta Values

I. Relative Risk When Disease Rate is 8/10,000

Alpha	Beta	Dis.Rate	Rel.Risk	Sample Size
0.05	0.05	8/10,000	8.9	2150
	0.10		7.5	
	0.15		6.7	
	0.20		6.0	
	0.25		5.5	
	0.30		5.1	
	0.35		4.7	
	0.40		4.3	
	0.45		4.0	
	0.49		3.8	
0.10	0.05	8/10,000	7.5	2150
	0.10		6.2	
	0.15		5.5	
	0.20		4.9	
	0.25		4.5	
	0.30		4.1	
	0.35		3.8	
	0.40		3.5	
	0.45		3.2	
	0.49		3.0	
0.15	0.05	8/10,000	6.7	2150
	0.10		5.5	
	0.15		4.8	
	0.20		4.3	
	0.25		3.9	
	0.30		3.6	
	0.35		3.2	
	0.40		3.0	
	0.45		2.7	
0.20	0.05	8/10,000	6.0	2150
	0.10		4.9	
	0.15		4.3	
	0.20		3.8	
	0.25		3.4	
	0.30		3.1	
	0.35		2.8	
	0.40		2.6	
	0.45		2.4	
0.25	0.05	8/10,000	5.5	2150
	0.10		4.5	
	0.15		3.9	
	0.20		3.4	
	0.25		3.1	
	0.30		2.8	
	0.35		2.5	
	0.40		2.3	
	0.45		2.1	
0.30	0.05	8/10,000	5.1	2150
	0.10		4.1	
	0.15		3.5	
	0.20		3.1	
	0.25		2.8	
	0.30		2.5	
	0.35		2.2	
	0.40		2.0	
	0.45		1.8	
0.33	0.05	8/10,000	4.9	2150
	0.10		3.9	
	0.15		3.4	
	0.20		2.9	
	0.25		2.6	
	0.30		2.4	
	0.35		2.1	
	0.40		1.9	
	0.45		1.7	

II. Relative Risk When Disease Rate is 8/100,000.

Alpha	Beta	Dis. Rate	Rel.Risk	Sample Size
0.05	0.05	8/100,000	65.9	2150
	0.10		52.6	
	0.15		44.8	
	0.20		38.8	
	0.25		34.1	
	0.30		30.4	
	0.35		26.9	
	0.40		23.8	
	0.45		21.2	
0.10	0.05	8/100,000	52.6	2150
	0.10		40.9	
	0.15		34.1	
	0.20		28.9	
	0.25		24.9	
	0.30		21.8	
	0.35		18.9	
	0.40		16.4	
	0.45		14.3	
0.15	0.05	8/100,000	44.8	2150
	0.10		34.1	
	0.15		28.0	
	0.20		23.4	
	0.25		19.8	
	0.30		17.1	
	0.35		14.5	
	0.40		12.4	
	0.45		10.6	
0.20	0.05	8/100,000	38.8	2150
	0.10		28.9	
	0.15		23.4	
	0.20		19.2	
	0.25		16.0	
	0.30		13.6	
	0.35		11.4	
	0.40		9.5	
0.33	0.05	8/100,000	28.2	2150
	0.10		20.0	
	0.15		15.5	
	0.20		12.2	
	0.25		9.8	
	0.30		8.0	
	0.35		6.4	
	0.40		5.1	
	0.45		4.1	

References

Cothern, Coniglio and Marcus (1986), "Estimating Risks to Health," *Environmental Science and Technology* III 20.

Cranor, C. (1983), "Epidemiology and Procedural Protections for Workplace Health in the Aftermath of the Benzene Case." *Industrial Relations Law Journal.*

_ _ _ _ _. (1987) "Some Public Policy Problems with Risk Assessment: How Good is the Use of the 95% Rule in Epidemiology?" Invited Symposium Paper, Pacific Division Meetings American Philosophical Association.

Feinstein, A. (1977), *Clinical Biostatistics.*

Freedman, D. and Zeisel, H. (forthcoming). "From Mouse to Man: The Quantitative Assessment of Cancer Risks," *Statistical Scienc*e.

Lee and Chang, A.C. (forthcoming), "An Evaluation of Transport Modeling of Organic Chemicals in Soil from Underground Fuel Tank Leaks."

Giere, R. (1981) *Understanding Scientific Reasoning.*

Mayo, D. (1989), "Towards a More Objective Understanding of the Evidence of Carcinogenic Risk." *PSA 1988*, Vol. 2, pp. 485-503..

National Cancer Institute, Demographic Analysis Section, Division of Cancer Cause and Prevention, (1981). Monograph No. 57. Surveillance, *Epidemiology and End Results: Incidence and Mortality Data* 1973-1977, pp. 662-63 and Table 51.

National Research Council, (1983), *Risk Assessment in the Federal Government: Managing the Process.* Washington, D.C.: U.S. Gov't Printing Office.

Needleman and Billinger (forthcoming), "Type II Fallacies in the Study of Childhood Exposure to Lead at Low Dose: A Critical and Quantitative Review."

Page, T. (forthcoming). *The Economics of Risk Assessment.*

Rescher, N. (1983), *Risk: A Philosophical Introduction to the Theory of Risk Evaluation and Management.* Washington, D.C.: University Press of America.

U.S. Congress, Office of Technology Assessment (1987), *Identifying and Regulating Carcinogens.* Washington, D.C.: U.S. Government Printing Office.

U.S. Congress, *Office of Technology Assessment* (1981),

Walter (1977), "Determination of Significant Relevant Risks and Optimal Sampling Procedures in Prospective and Retrospective Studies of Various Sizes." *American Journal of Epidemiology* 105: 387-91.